JOURNAL OF DATA SCIENCE | Vol.2024:60
eISSN:2805-5160

Text to Image Generation Using Machine Learning

Rishab Tiwari[1], Chitra K[2]

[1,2]Dayananda Sagar Academy of Technology and Management, Bangalore 560082 India

Email: rishabtiw045@gmail.com[1], chitra-mca@dsatm.edu.in[2]

Abstract

A method called text-to-image involves creating images automatically from provided written descriptions. It contributes significantly to artificial intelligence by tackling the problem of integrating textual and visual input. One of the usefulness of automatic picture synthesis is the generation of images using conditional generative models. For this, Generative Adversarial Networks (GANs) are frequently employed. Using GANs, recent developments in the sector have made significant progress. An outstanding illustration of deep learning's potential is the transformation of text into images. It is difficult to create a text-to-image synthesis system that consistently creates realistic graphics based on predetermined criteria. Many of the existing algorithms in this field struggle to produce visuals that precisely match the given text. In order to solve this issue, we carried out a research work where we concentrated on developing the generative adversarial network (GAN), a deep learning-based architecture. The aim of this research work is to create a system that allows you to generate images that are semantically consistent.

Keywords

Generative Adversarial Networks, Convolutional neural network, deep learning

Introduction

The goal of text-to-image (T2I) creation is to produce aesthetically accurate and semantically consistent images from textual descriptions. Given its significance in numerous applications, including as photo editing, art creation, and computer-aided design, this particular problem has attracted a great deal of focus in the deep learning community. The large dimensionality of the output space and the semantic gap between the textual and visual domains are the key reasons why it also poses substantial obstacles.

Once the technology is ready for commercial usage, the creation of graphics from natural language has enormous promise for a variety of future applications. As generative models, Generative Adversarial Networks (GANs) are capable of producing new content. The goal of text-to-image synthesis is to turn verbal descriptions into aesthetically pleasing pictures. GAN models are now frequently employed in this industry to get better outcomes. The fact that a single text description might take on various configurations presents a barrier for deep learning.

Submission: 23 October 2024; **Acceptance:** 25 November 2024

By giving the model the proper training, this problem can be solved. As generative adversarial networks (GANs) have developed, they have proven to be remarkably effective at a variety of tasks involving images, such as picture synthesis, image super-resolution, data augmentation, and image-to-image conversion.

Methodology

Generative Adversarial Network (GAN), which consists of a generator and a discriminator, is the deep learning method we utilized. For text to picture generation, we also employed Tensorflow, Numpy, NLTK, and Tensorlayer. Tensorflow is essentially a machine learning library. In comparison to other deep learning libraries, it compiles quicker. Additionally, both CPU and GPU computing units are supported. We use the Python Pickle module for data serialization in our network design. By converting objects into byte streams, this module enables us to easily store the data in files and transfer it between different systems and applications.

Generative Adversarial Networks. is an unsupervised learning strategy that trains a generative model to produce fresh examples. GANs can be used in many domains, such as the synthesis of images and sounds, and use neural networks to create new instances of data. The term "generative" is used to describe learning a model that can produce fresh data in the context of GANs, and the model is trained using neural networks [6]. The discriminator and the generator are the two component sections of the GAN.

Generator:
The generator in the Generative Adversarial Network (GAN) is in charge of producing fresh instances of data, which are frequently bogus or synthetic examples. The discriminator, whose job it is to discern between actual and fraudulent data, is then shown these created samples. Goal of the generator is to provide samples that successfully trick or perplex the discriminator, making it challenging for the hater of diversity to correctly determine whether a sample is real or artificial. The generator and discriminator's competitive procedure encourages learning and long-term development of both models.

Discriminator:
The discriminator in a Generative Adversarial Network (GAN) is in charge of separating real samples from phony samples produced by the generator. Deep neural networks are used for the generator and discriminator. Goal of the generator is to is to trick the discriminator by creating phony instances that resemble actual data. The discriminator, on the other hand, seeks to accurately recognize and categorize genuine data samples. As a result, there is competition between the generator and discriminator.

Proposed Methodology:
We used Conditional Generative Adversarial Networks (GANs) in combination with Recurrent Neural Networks (RNNs) and Convolutional Neural Networks (CNNs) as part of the training phase of our deep learning-based generative models to produce meaningful images based on textual descriptions. Our data set included floral photos and the language descriptions that go with them.

We preprocessed the textual data and scaled the photos to a fixed dimension in order to generate convincing visuals from text using GANs. We parsed the dataset's caption sentences, built a vocabulary list, and gave each caption a special ID. The photos were loaded and appropriately scaled. These previously processed textual and visual data then served as the foundation for our suggested model.

We used RNNs to extract the contextual information from the text sequences. The association between words at various time stamps was created via RNNs, which allowed the model to comprehend the textual descriptions. We combined RNN and CNN to carry out the texttoimage mapping. Without human input, CNN retrieved pertinent details from the photographs.

An input series of textual descriptions was fed into the RNN during training, and the RNN transformed the text into 256-word embeddings. Then, a 512-dimensional noise vector was concatenated with these word embeddings. With a gated-feedback generator of size 128 and a batch size of 64, we trained our model while feeding the generator inputs of noise and text.

We used the textual description's extracted semantic data as input for the generator model. The generator used this semantic data to translate the distinguishing details into pixel-level data and produce related images. The discriminator then used these generated images as input along with accurate or inaccurate verbal descriptions and authentic sample images from the collection.

The model receives as input during training a series of unique pairings made up of photos and their matching textual descriptions. This is done in order to achieve the discriminator's goals. The input pairings consist of produced images with accurate text descriptions, erroneous images with inaccurate text descriptions, and genuine photographs with accurate text descriptions.

To enable the model to ascertain whether a specific image and text pairing are in alignment with one another, actual image and real text combinations are used. A mismatch between the image and the caption is indicated if an incorrect image is matched with a true written description.

The discriminator is taught to distinguish between authentic and artificial images. The discriminator's classification performance is initially mainly concerned with telling correct images from false ones. In order to enhance weight updates and give training feedback to both the generator and discriminator models, the loss is estimated during training.

Results and Discussion

The user enters text into the GUI application, which then processes it and displays the accompanying image. As a result, the application displays an image that matches the written description.

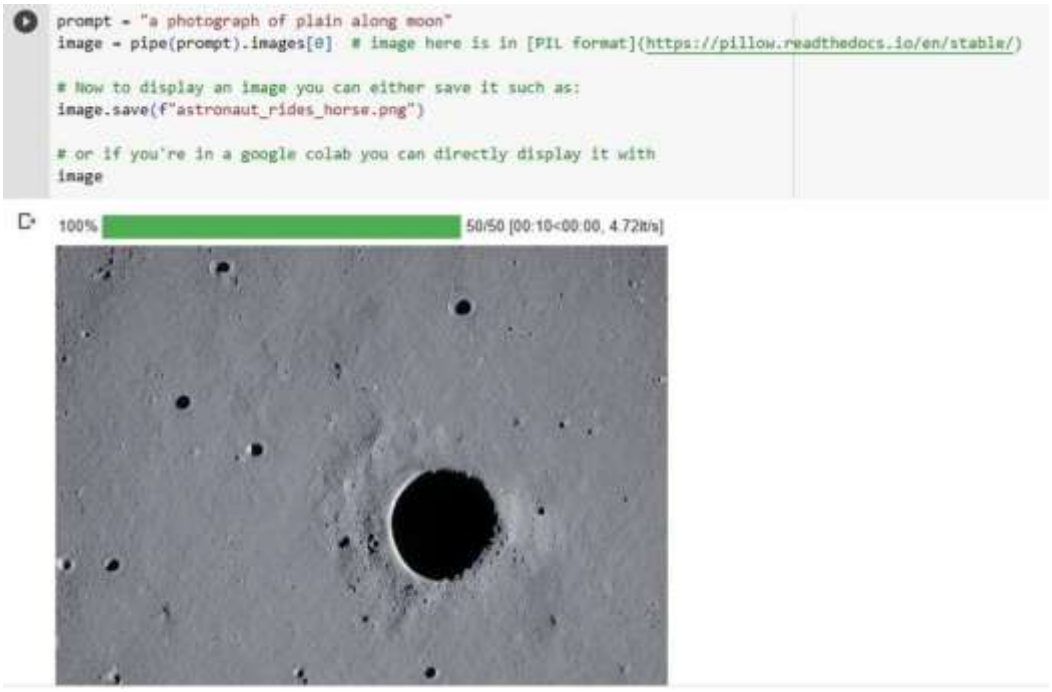

Figure 1 First Image Generation

The proposed GUI application represents a significant advancement in the field of text-to-image generation, showcasing its ability to seamlessly process textual input and produce corresponding visual outputs. By allowing users to input text descriptions, the application leverages advanced generative adversarial network (GAN) models to interpret the semantics of the text and translate them into visually meaningful images. This process highlights the robustness of the underlying architecture in handling complex natural language inputs and converting them into accurate graphical representations.

Figures 1 and 2 illustrate the successful implementation of the application's core functionality. In Figure 1, the generated image accurately aligns with the description provided by the user, showcasing the application's ability to synthesize coherent and realistic visuals. Similarly, Figure 2 demonstrates the application's consistency in maintaining semantic relevance across varied textual inputs. These results affirm the system's potential in scenarios where automated image generation based on descriptive input is essential, such as creative content development, visual storytelling, and educational tools.

The primary strength of the system lies in its interactive nature, which enhances user engagement by providing immediate visual feedback. Additionally, the ability to generate semantically accurate images underscores the effectiveness of the deep learning architecture employed, particularly in addressing challenges like maintaining semantic consistency and realism in generated visuals. However, some limitations may persist, such as handling ambiguous or overly complex text inputs, which can impact the quality or relevance of the generated image. Further optimization of the system's text interpretation capabilities and image synthesis processes could address these challenges.

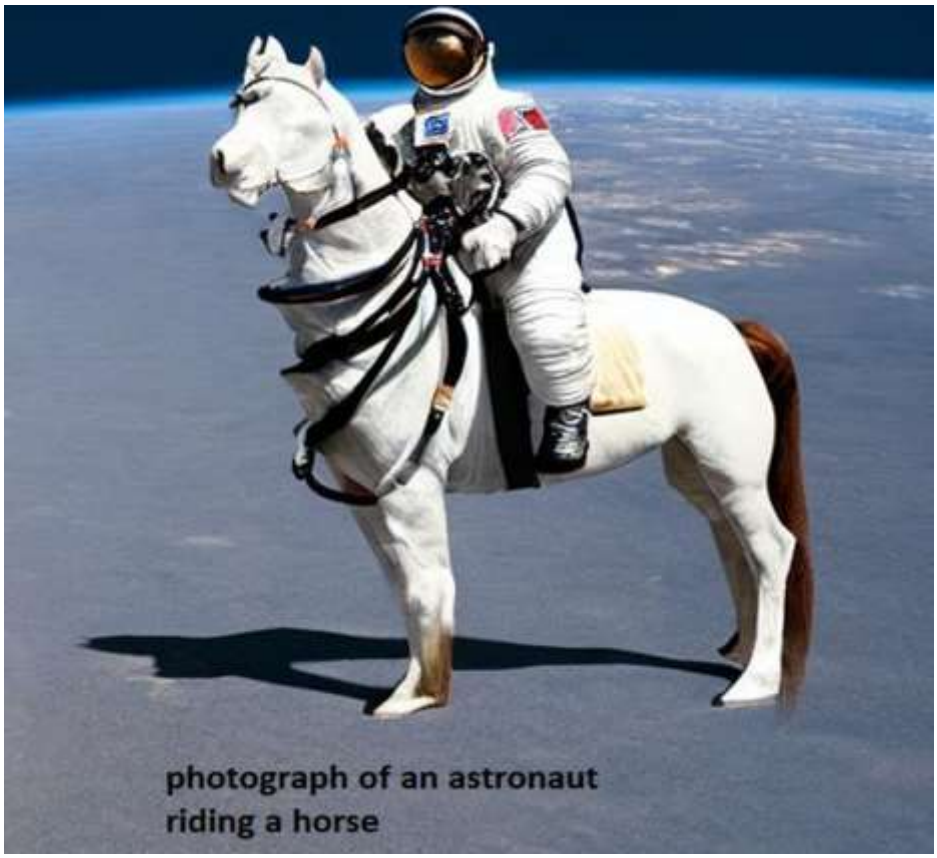

Figure 2 Second Image Generation

Conclusion

This research highlights the potential of generative adversarial networks (GANs) in bridging the gap between textual descriptions and visual representation. By addressing the limitations of existing algorithms in producing semantically consistent and realistic images, the proposed deep learning-based architecture demonstrates significant advancements in text-to-image synthesis. This study not only showcases the transformative capabilities of deep learning but also lays the groundwork for future innovations in integrating textual and visual modalities. The resulting system represents a meaningful step forward in creating AI-driven solutions that generate images directly aligned with given textual descriptions, contributing to advancements in artificial intelligence and creative automation.

References

Akanksha Singh 1, Sonam Anekar 2, Ritika Shenoy 3, Prof. Sainath Patil 4,(2021). Text to Image using Deep Learning. International Journal of Engineering Research & Technology (IJERT). ISSN: 2278-0181. Vol. 10 Issue 04, April-2021 https://www.ijert.org/research/text-to-image-using-deep-learning-IJERTV10IS040132.pdf

Bodnar, C. (2018). Text to Image Synthesis Using Generative Adversarial Networks. *ArXiv (Cornell University)*. https://doi.org/10.48550/arxiv.1805.00676

Brock, A., Donahue, J., & Simonyan, K. (2018). Large Scale GAN Training for High Fidelity Natural Image Synthesis. *ArXiv (Cornell University)*. https://doi.org/10.48550/arxiv.1809.11096

H. Zhang *et al*. (2017), "StackGAN: Text to Photo-Realistic Image Synthesis with Stacked Generative Adversarial Networks," *2017 IEEE International Conference on Computer Vision (ICCV)*, Venice, Italy, pp. 5908-5916, https://doi.org/10.1109/ICCV.2017.629

Jeon, E., Kim, K., & Kim, D. (2021). FA-GAN: Feature-Aware GAN for Text to Image Synthesis. *arXiv preprint arXiv:2109.00907*. https://arxiv.org/abs/2109.00907

Kataoka, Y., Matsubara, T., & Uehara, K. (2016). Image generation using generative adversarial networks and attention mechanism. *Annual ACIS International Conference on Computer and Information Science*. https://doi.org/10.1109/icis.2016.7550880

Li, J., Liu, X., & Zheng, L. (2023). Factor Decomposed Generative Adversarial Networks for Text-to-Image Synthesis. *arXiv preprint arXiv:2303.13821*. https://arxiv.org/abs/2303.13821

Reed, S., Akata, Z., Yan, X., Logeswaran, L., Schiele, B., & Lee, H. (2016). Generative Adversarial Text to Image Synthesis. *Proceedings of the 33rd International Conference on Machine Learning (ICML)*, 1060–1069. https://arxiv.org/abs/1605.05396

Tao, M., Tang, H., Wu, F., Jing, X.-Y., Bao, B.-K., & Xu, C. (2020). DF-GAN: A Simple and Effective Baseline for Text-to-Image Synthesis. *arXiv preprint arXiv:2008.05865*. https://arxiv.org/abs/2008.05865

Tibebu, H., Malik, A., & De Silva, V. (2022). Text to Image Synthesis using Stacked Conditional Variational Autoencoders and Conditional Generative Adversarial Networks. *arXiv preprint arXiv:2207.03332*. https://arxiv.org/abs/2207.03332